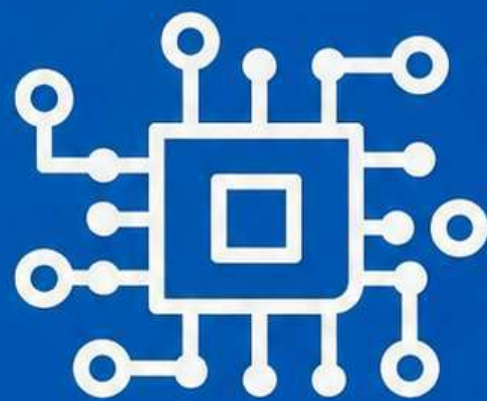




TECH SCIENCE

ISSN 3030-3702

**TEXNIKA FANLARINING
DOLZARB MASALALARI**

**TOPICAL ISSUES OF TECHNICAL
SCIENCES**



**№ 4-maxsus
son (4) 2026**

BOSH MUHARRIR:

ABBAZOV ILHOM ZAPIROVICH

TAHRIR HAY'ATI:

Usmankulov Alisher Kadirkulovich — Texnika fanlari doktori, professor, Jizzax, O'zbekiston;

Fayziyev Xomitxon – texnika fanlari doktori, professor, Toshkent, O'zbekiston;

Rashidov Yusuf Karimovich – texnika fanlari doktori, professor, Toshkent, O'zbekiston;

Adizov Bobirjon Zamirovich – Texnika fanlari doktori, professor, Toshkent, O'zbekiston;

Abdunazarov Jamshid Nurmuxamatovich — Texnika fanlari doktori, dotsent, Jizzax, O'zbekiston;

Eshonkulov Ravshan Abdurazakovich — Texnika fanlari doktori, dotsent, Qashqadaryo, O'zbekiston;

Umarov Shavkat Isomiddinovich – Texnika fanlari doktori, professor, Jizzax, O'zbekiston;

Bozorov G'ayrat Rashidovich – Texnika fanlari doktori, Buxoro, O'zbekiston;

Maxmudov Muxtor Jamolovich – Texnika fanlari doktori, Buxoro, O'zbekiston;

Asatov Nurmuxammat Abdunazarovich – Texnika fanlari nomzodi, professor, Jizzax, O'zbekiston;

Mamayev G'ulom Ibroximovich – Texnika fanlari bo'yicha falsafa doktori (PhD), Jizzax, O'zbekiston;

Ochilov Abduraxim Abdurasulovich – Texnika fanlari bo'yicha falsafa doktori (PhD), Buxoro, O'zbekiston.

OAK Ro'yxati

Mazkur jurnal O'zbekiston Respublikasi Oliy ta'lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestatsiya komissiyasi Rayosatining 2025-yil 8-maydagi 370-son qarori bilan texnika fanlari bo'yicha ilmiy darajalar yuzasidan dissertatsiyalar asosiy natijalarini chop etish tavsiya etilgan ilmiy nashrlar ro'yxatiga kiritilgan.

Muassislar: "SCIENCEPROBLEMS TEAM" mas'uliyati cheklangan jamiyati;
Jizzax politexnika instituti.

TECHSCIENCE.UZ- TEXNIKA**FANLARINING DOLZARB****MASALALARI** elektron jurnali

15.09.2023-yilda 130343-sonli

guvohnoma bilan davlat ro'yxatidan o'tkazilgan.

TAHRIRIYAT MANZILI:

Toshkent shahri, Yakkasaroy tumani, Kichik Beshyog'och ko'chasi, 70/10-uy.

Elektron manzil:

scienceproblems.uz@gmail.com

Barcha huquqlar himoyalangan.

© Scienceproblems team, 2026-yil

© Mualliflar jamoasi, 2026-yil

MUNDARIJA

<i>Mamirov Uktam, Khudoyberdiyev Asror</i> FUZZY-LOGIC-BASED ASSESSMENT OF THE DURABILITY CRITERION OF A CONTINUOUS CASTING MACHINE	4-12
<i>Shanning Fu, Ya Han, Bekmurotov Dilshod</i> CONTROLLED EVALUATION OF STATE-AWARE AND VERIFIABLE RAG FOR ENGLISH-LANGUAGE TUTORING: RETRIEVAL RELIABILITY, GROUNDING, AND VERIFICATION	13-26
<i>Fayzullayev Xayriddin</i> MUHANDISLIK GRAFIKASI FANINI O'RGANISHDA "TEKISLIK LARNING KESISHISHI" MAVZUSINI O'QITISH USULI	27-32
<i>Muinov Ulug'bek, Usmonov Botir</i> KICHIK QUUVATLI SHAMOL ENERGETIK TIZIMIDA PI-ROSTLAGICH ASOSIDA KUCHLANISHNI BOSHQARISH	33-44
<i>Sayitov Shavkatjon, Sarimsakov Olimjon</i> PAXTA XOMASHYOSINI QURITISHNING KO'P PARAMETRLI AVTOMATLASHTIRILGAN MONITOTING TIZIMINI ISHLAB CHIQISH	45-50
<i>Haitov Bozar</i> BISMUT VOLFRAMAT (Bi_2WO_6 , BiWO_4) NI SINTEZ QILISH USULLARI: ZAMONAVIY YONDASHUVLAR TAHLILI	51-58
<i>Ishnazarov Oybek, Hoshimov Ural</i> KOMPRESSOR USKUNASINI AVTOMATLASHTIRILGAN ISHLASH REJIMIDAGI DINAMIK JARAYONLARNING TAHLILI	59-67
<i>Shavazov Abdulatif, Turaxanov Sherzod, Ilekeroval Indira</i> BOSHQARILUVCHI ELEKTR YURITMALAR ISH REJIMLARINI OPTIMALLASHTIRISH USULLARINI TAKOMILLASHTIRISH.....	68-75
<i>Nurlibayeva Zarina</i> TIKUV MASHINASI IGNA IPINI TARANGLAGICHINING QAYISHQOQ ELEMENTLI KONSTRUKTIV PARAMETRLARINI TADBIQ ETISH	76-83
<i>Bazarov Baxtiyor, Odilov Odiljon, Bo'tayev Azizjon</i> SINTETIK GAZ YONILG'ISIDAN FOYDALANIB CHEVROLET COBALT VA CHEVROLET TRACKER YENGIL AVTOMOBILLARINING EKSPLOATATSION KO'RSATKICHLARINI OSHIRISH: MATLAB/SIMULINK ASOSIDAGI MODELLASHTIRISH TADQIQOTI	84-90
<i>Karimxodjayev Nazirjon, Ismatov Biloldin</i> ELEKTROMOBIL BATAREYALARINI ZARYADLASH JARAYONIDA YASHIRIN KRITIK HOLATLARNI AMALIY DIAGNOSTIKA QILISH METODIKASI	91-99
<i>Musurmonov Ulug'bek</i> POLIMER MODIFIKATSIYALANGAN BITUM ASOSIDA ASFALT-BETON QORISHMASINING FIZIK-MEXANIK XUSUSIYATLARINI TADQIQ ETISH.....	100-107

CONTROLLED EVALUATION OF STATE-AWARE AND VERIFIABLE RAG FOR ENGLISH-LANGUAGE TUTORING: RETRIEVAL RELIABILITY, GROUNDING, AND VERIFICATION

Shanning Fu

National Research University "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers"
Doctoral Student
ORCID: 0009-0007-3280-7991
Tashkent, Uzbekistan

Ya Han

Shandong Hongjitang Pharmaceutical Group Co., Ltd., Jinan, China
Research Staff, Bachelor of Science
ORCID: 0009-0007-0384-5745
Email: 15853899107@163.com

Dilshod Bekmurotov

National Research University "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers"
PhD, Senior Lecturer
ORCID: 0000-0001-8306-1500
Tashkent, Uzbekistan

Annotation. We evaluated a state-aware, verifiable RAG tutor on 72 English-language tasks across three conditions and two runs. Linguistic accuracy showed no condition effect, while retrieval faithfulness tracked evidence access, especially under weak T3 retrieval. Added learner-state and verification controls did not yield reproducible gains; human validation supported separate evaluation.

Key words: retrieval-augmented generation, English-language tutoring, learner modeling, verification, retrieval faithfulness, LLM-as-a-Judge, human validation.

INGLIZ TILINI O'RGATISH UCHUN O'RGANUVCHI HOLATINI HISOBGA OLUVCHI VA VERIFIKATSIYALANADIGAN RAGNI NAZORATLI BAHOLASH: QIDIRUV ISHONCHLILIGI, DALILLARGA ASOSLANGANLIK VA VERIFIKATSIYA

Shanning Fu

"Toshkent irrigatsiya va qishloq xo'jaligini mexanizatsiyalash muhandislari instituti"
milliy tadqiqot universiteti
Doktorant
Toshkent, O'zbekiston

Ya Xan

"Shandong Hongjitang Pharmaceutical Group Co., Ltd.", Tszinan, Xitoy

Tadqiqotchi xodim, Tabiiy fanlar bakalavri

Dilshod Bekmuratov

"Toshkent irrigatsiya va qishloq xo'jaligini mexanizatsiyalash muhandislari instituti" milliy tadqiqot universiteti
PhD, katta o'qituvchi

Annotatsiya. Biz holatni hisobga oluvchi va verifikatsiyalanadigan RAG repetitorini 72 ta ingliz tili topshirig'ida, uch sharoit va ikki ishga tushirishda baholadik. Lingvistik aniqlik sharoitlar bo'yicha farq qilmadi; qidiruv dalillariga sodiqlik, ayniqsa T3 zaif qidiruvda, dalillarga kirish bilan bog'liq bo'ldi. Qo'shimcha mexanizmlar barqaror ustunlik bermadi; inson validatsiyasi alohida baholashni qo'llab-quvvatladi.

Kalit so'zlar: qidiruv bilan boyitilgan generatsiya, ingliz tilidan repetitorlik, o'quvchi modeli, tekshirish, qidiruv natijalariga sodiqlik, hakam sifatida LLM, inson tomonidan tasdiqlash.

DOI: <https://doi.org/10.47390/ts-v4is/4y2026N02>

1. Introduction

Generative artificial intelligence enables digital learning systems to provide explanations, examples, feedback, and conversational support. In language education, LLMs support writing, feedback, content generation, and tutoring, but fluency alone does not establish reliability or pedagogical value [9,26,27]. Reliable tutoring therefore requires inspectable evidence, intended state sensitivity, and credible quality control.

LLM tutoring extends adaptive and intelligent tutoring traditions, where learner models support individualized instruction [14,22,24]. In prompt-based systems, however, appending learner information does not by itself show that responses are meaningfully learner-sensitive.

RAG grounds generation in external evidence [10], but reliability depends on query formulation, retrieval quality, evidence selection, and evidence use [25]. Unsupported claims and imperfect retrieval remain possible [4,17,19], so answer quality and grounding must be evaluated separately.

Verification adds control but does not guarantee self-correction. Iterative refinement and tool-supported critique can improve outputs [5,15], yet reliable external feedback is often critical [8]. A same-family verifier given the same deficient evidence may therefore have little new information for repair.

LLM judges scale free-form evaluation [12], but bias and construct-specific disagreement with humans are documented [2,23]. Because correctness, feedback quality, learner uptake, and pedagogical effectiveness are distinct [6,13,21], we validate automated metrics separately against linguistic experts rather than treating the judge as a gold standard.

Recent controlled educational RAG studies include Adenuga's same-model comparison of prompt-only and learner-state-aware RAG tutoring, KITE's multimodal RAG with intent-aware Socratic tutoring, and PRAG-EDU's grade-level profiles with retrieved course materials [1,7,16]. They confirm that educational RAG and learner-aware conditioning are active evaluation targets. A narrower gap remains: the incremental reliability of retrieval, learner-state conditioning, and verification is unclear when the generation backbone is fixed, a retrieval-only intermediate condition is retained, B1 and B2 receive identical evidence, and automated metrics are calibrated separately against linguistic experts. We address this gap through controlled subsystem-reliability analysis.

The objective is to diagnose a lightweight SAVR configuration under controlled comparison, not to assume architectural superiority. We ask whether (RQ1) the three conditions differ in Linguistic Accuracy; (RQ2) retrieval quality is associated with Retrieval Faithfulness; (RQ3) added B2 mechanisms produce reproducible improvement across primary automated outcomes under the prespecified hierarchy when B1/B2 retrieval is identical; and (RQ4) automated metrics agree sufficiently with independent experts for inferential use. Because B2 bundles learner-state conditioning and verification, adaptation is not treated as a separately identified causal effect.

The study provides a controlled reliability framework for RAG tutoring by comparing B0, retrieval-only B1, and B2 with identical B1 evidence plus learner state and verification. This design separates evidence access from downstream control, while cross-run audits, retrieval diagnostics, and human validation distinguish linguistic accuracy from evidence faithfulness and identify metrics suitable for primary inference.

2. Literature Analysis and Methodology

2.1 Literature Analysis

RAG reliability spans evidence access and evidence-faithful generation, so evaluation should separate retrieval/context quality from answer faithfulness and overall quality [3,20]. A strong LLM may answer correctly from parametric memory despite irrelevant retrieval, producing acceptable language but weak grounding. Context-faithfulness studies also show competition between parametric memory and external evidence [11,28]. We therefore evaluate retrieval performance and Retrieval Faithfulness separately.

Learner-state modeling serves a different function. Learner models support individualized instruction [14,24], but a learner-state field alone does not demonstrate adaptation. Learner sensitivity should appear in explanation depth, error focus, example choice, or feedback sequencing. Retrieval therefore determines what evidence is available, whereas learner state concerns how that evidence is presented.

Verification is treated as an empirical control, not an assumed safeguard. Automated-correction research indicates that correction is more dependable when the verifier receives useful information unavailable to the generator [8,18]. A same-backbone verifier sharing weak evidence may therefore reproduce rather than repair the generator's failure.

We also treat LLM-based evaluation as a construct-validation problem. A judge may align with experts on one metric but fail on another or conflate weak retrieval support with factual error [2,12,23]. Each automated metric is therefore validated separately against two linguistic experts, keeping retrieval, learner-state conditioning, verification, and evaluation independently auditable.

2.2 Materials and Methodology

The experiment used a paired controlled comparison with three conditions. B0 received only the tutoring prompt; B1 used the same LLM plus the top three cards from a deterministic local retriever; B2 received the same retrieved evidence and added learner-state information for T3 tasks plus a post-generation verifier. Because learner state and verification were bundled in B2, their independent causal effects were not estimated (Table 1).

The benchmark contained 72 English-language tutoring tasks: 24 target linguistic phenomena crossed with T1 Diagnosis and Correction, T2 Explanation and Usage, and T3 Stateful Adaptive Tutoring. It included 36 operational A2 and 36 operational B1 tasks across six linguistic domains, with six no-error T1 controls for unnecessary correction. The A2/B1

labels standardized task difficulty and response expectations and were not externally certified CEFR test items.

Each task record included the prompt, target phenomenon, operational band, gold target, acceptable alternatives, prohibited corrections, gold rule, feedback guidance, optional learner state, source URL, and primary metrics. The study-specific benchmark was designed for controlled coverage rather than external certification or population inference. Tasks and all gold-standard constraints were frozen before formal generation and were not revised after outputs were observed.

The retrieval corpus contained 24 compact knowledge cards, one per target phenomenon. A local TF-IDF retriever with unigram and bigram features ranked cards by cosine similarity using the task prompt alone; learner-state records, gold answers, target labels, SLA tags, and other answer-bearing metadata were excluded to prevent leakage. B1 and B2 received the same top-three cards and retrieval scores for each task, holding evidence access fixed. Post hoc diagnostics used gold-card rank, top-ranked similarity, the Top-1-minus-Top-2 similarity gap, and a gold-card retrieval margin defined as $\text{similarity}(\text{gold}) - \text{similarity}(\text{best non-gold})$. Because the archive retained scores only for the top three cards, exact gold-card margins were available only when the gold card was retrieved in the top three; otherwise only an upper bound was possible, so no exact all-task margin test was performed.

Generation used DeepSeek V4-Flash through the DeepSeek API with thinking disabled, temperature = 0.2, top_p = 1.0, and max_tokens = 300; responses targeted approximately 120 words or fewer. B0 used the task prompt alone, B1 added the top-three retrieved cards, and B2 added the same cards plus the standardized learner-state record for T3 tasks. B2 candidates were then checked by a verifier from the same model family for unnecessary correction, materially incorrect rules, unsupported evidence use, learner-history misuse, and clearly inappropriate pedagogical feedback, returning ACCEPT or REVISE. Verifier contribution was assessed from observed interventions and candidate-to-final text changes, not the decision label alone. Candidate/final responses, decisions, identifiers, token usage, and raw API records were retained; the process audit used exact and normalized text comparisons to distinguish substantive from formatting-only changes.

Two formal runs used the same frozen inputs and protocol: Run 1 (1 September 2026) was the primary analysis and Run 2 (6 September 2026) an operational robustness replication. Each run contained 72 tasks x 3 conditions = 216 episodes plus 72 B2 verifier calls. Tasks, knowledge base, prompts, model endpoints, generation parameters, retrieval, and scoring were held constant. The environment used Python 3.14.5, requests 2.34.2, and scikit-learn 1.8.0. Structural audits checked completeness, duplicates, API integrity, non-empty outputs, returned model names, B1/B2 retrieval identity, and raw-file completeness; hashes were retained to detect protocol drift. Runs were analyzed separately rather than pooled as independent observations. Figure 1 summarizes the workflow.

Reproducibility identifiers. Run 2 archives preserve provider/model identifiers, system fingerprint, UTC execution ranges, request-specific SHA-256 prompt hashes, and frozen task/knowledge-base hashes; Judge V1.3 identifiers are also retained. Supplementary Table S2 reports the full record. Run 1 per-case raw payloads are absent, so unverified identifiers or hashes were not reconstructed. Credentials are excluded.

Automated evaluation used a separate DeepSeek V4-Pro evaluator (Judge V1.3) with six 0-2 dimensions: Linguistic Accuracy, Correction Restraint, Retrieval Faithfulness, Unsupported

Content, Pedagogical Appropriateness, and Adaptivity. The judge ran separately from generation and, for B1/B2, received the actual retrieved evidence so Retrieval Faithfulness was scored against evidence available to the tutor; that metric applied only when a material evidence-bearing claim was present. Linguistic Accuracy and Retrieval Faithfulness were primary outcomes; Correction Restraint and Pedagogical Appropriateness were secondary; Unsupported Content and Adaptivity were exploratory after human validation identified construct or ceiling limitations. Judge V1.3 was not retuned after human labels were observed.

Human validation used 18 stratified tasks spanning six linguistic domains and all three task types, yielding 54 anonymized condition outputs. Two linguistic experts independently applied the same 0-2 rubric with system labels and automated scores hidden. They received the task information required for each construct, including retrieved evidence for Retrieval Faithfulness. Exact agreement, within-one agreement, and quadratic weighted kappa were computed where estimable, and original ratings were preserved. Because the archive does not document rater rank, assessment experience, or noninvolvement in SAVR development, these are not claimed as validity evidence.

Within each run, outcomes applicable to all three conditions were compared with Friedman tests; significant omnibus effects would be followed by paired Wilcoxon signed-rank tests with Holm correction, with Kendall's W as the Friedman effect-size index. Retrieval Faithfulness was analyzed pairwise for B1 versus B2 because B0 had no retrieved evidence. Because paired ordinal scores were tie-dominated, the primary summary was the number and direction of non-zero changes with two-sided exact sign tests; Wilcoxon tests were sensitivity analyses. Spearman correlations quantified the association between gold-card rank and Retrieval Faithfulness. Two-sided $\alpha = .05$ was used, while interpretation emphasized effect magnitude, tie structure, human-validation evidence, and replication. The 72-task benchmark was coverage-driven rather than power-based, and no equivalence margin was prespecified.

Task-type heterogeneity in gold-card rank, top-ranked cosine similarity, and the Top-1-minus-Top-2 gap was tested across the 24 matched linguistic phenomena with Friedman tests. Because the diagnostic targeted the T3 deficit, planned paired contrasts of T3 versus T1 and T3 versus T2 used Wilcoxon tests with Holm correction within each metric; exact gold-card margins among Top-3-hit tasks were descriptive because conditioning on retrieval success changes the estimand. Run-to-run stability used exact agreement, within-one-category agreement, and Spearman correlation on matched Judge V1.3 scores. Linguistic Accuracy included all 216 matched episodes; Retrieval Faithfulness included 142 matched scorable RAG episodes because two Run 2 B1 outputs contained no material evidence-bearing claim. With ceiling- and tie-limited distributions, exact and within-one agreement were primary reproducibility descriptors, and rank correlation was interpreted only when score variation was sufficient. The study used synthetic controlled tutoring tasks and AI-generated outputs; no learners, learner records, or personal learning outcomes were collected.

Table 1. Experimental conditions.

Condition	Base model	Top-3 RAG	Learner state	Verifier
B0	DeepSeek V4-Flash	No	No	No
B1	DeepSeek V4-Flash	Yes	No	No

Condition	Base model	Top-3 RAG	Learner state	Verifier
B2	DeepSeek V4-Flash	Same as B1	T3 only	Yes

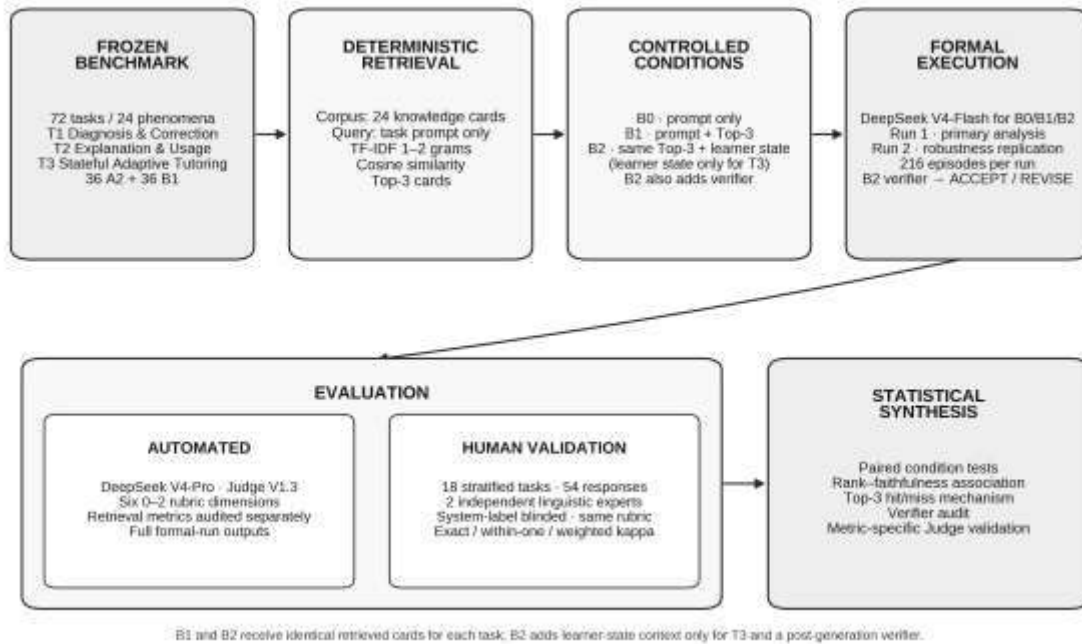


Figure 1. SAVR tutoring architecture and controlled evaluation workflow. B0 used the task prompt only; B1 added deterministic Top-3 TF-IDF retrieval from 24 frozen knowledge cards; B2 added the same retrieval plus learner-state context for T3 and a post-generation verifier. Run 1 was the primary analysis and Run 2 a robustness replication (216 episodes each). DeepSeek V4-Pro Judge V1.3 scored all formal outputs, and two linguistic experts independently scored a stratified 54-response subset.

3. Discussion

The results do not support a simple 'more components are better' interpretation. DeepSeek V4-Flash already produced near-ceiling linguistic correctness, leaving little observed headroom for RAG or verification on the primary accuracy score. Because the benchmark was coverage-driven, no equivalence margin was prespecified, and distributions showed ceiling effects and ties, non-significant tests do not establish equivalence. The narrower conclusion is that no reproducible architecture-level advantage was detected under the tested configuration.

The two-run design supports task-level reproducibility. Linguistic Accuracy showed very high exact agreement despite little rank variance, while Retrieval Faithfulness retained high exact agreement and strong rank stability. Thus retrieval-linked grounding behavior was reproducible under the frozen client-side protocol. Because hosted APIs were used, Run 2 is an operational rather than bitwise backend replication.

The most stable result was the retrieval-faithfulness association: retrieval varied sharply by task family, and gold-card rank was strongly associated with Retrieval Faithfulness. This supports decomposed RAG evaluation, as in RAGAs and ARES [3,20], where evidence access and answer faithfulness are audited separately. Because faithfulness is defined relative to supplied evidence, the association is partly structural; our contribution is its magnitude, task-family heterogeneity, and cross-run replication, not a causal claim about overall response quality.

T3 made the upstream constraint concrete. Retrieval used the task prompt only, while learner state was supplied separately to B2. T3 did not simply show lower cosine similarity than T1; the correct card was more often outranked, and its median gold-versus-best-non-gold margin among Top-3 hits was slightly negative. This is consistent with weaker lexical discrimination in stateful-adaptive prompts, but Top-3-only archives and the absence of query/card ablations preclude a unique causal diagnosis. A direct next test is an evidence-oriented retrieval query separated from learner state.

The state-plus-verification bundle did not reliably compensate for weak evidence access. B2 showed no stable Retrieval Faithfulness advantage over B1, and the verifier rarely changed final responses. This accords with self-correction research showing that verification is most informative when it receives useful signals unavailable to the generator [8,18]. In the same-family, same-evidence configuration tested here, missing evidence limited corrective opportunity.

Human validation narrowed the automated metrics. Linguistic Accuracy and Retrieval Faithfulness had the strongest support, whereas Unsupported Content and Adaptivity did not; Correction Restraint and Pedagogical Appropriateness were ceiling- or zero-variance limited. The judge sometimes conflated weak grounding with unsupported factuality and assigned maximum Adaptivity more often than experts. Same-provider generator, verifier, and judge components may share biases, so these findings support construct-specific validation rather than accepting one LLM judge globally [2,12,23].

Key limitations are the coverage-driven 72-task benchmark without an equivalence margin; the nonstandardized 24-card corpus; bundled learner-state and verification effects; related generator, verifier, and judge model families; hosted-endpoint rather than bitwise replication; evidence-visible faithfulness ratings; no real learners or learning outcomes; and Top-3-only similarity archives.

Future work should test dense, hybrid, and reranked retrieval; evidence-oriented queries; larger ambiguous corpora; separate learner-state and verifier ablations; independent verification; broader human validation; standardized benchmarks; cross-language replication; and predefined equivalence thresholds.

4. Results

Both formal runs passed structural audit: all 216 task-condition episodes were present in each run, with no duplicate pairs, API errors, empty final responses, or B1/B2 retrieval mismatches. The frozen inputs, retrieval implementation, and scoring protocol were unchanged, and deterministic retrieval was identical across runs.

Overall retrieval achieved Recall@1=.472, Recall@3=.625, and MRR=.576; the gold card fell outside the top three for 27 of 72 tasks (37.5%). T1 achieved .583/.708/.674 for Recall@1/Recall@3/MRR, T2 .667/.792/.743, and T3 only .167/.375/.312. Thus, T3 had the weakest evidence access before learner-state or verification processing (Table 2; Figure 2).

A post hoc diagnostic clarified the T3 failure mode. Median gold-card rank was 1 for T1, 1 for T2, and 6 for T3; the matched difference was significant (Friedman chi-square(2)=10.784, $p=.0046$), and T3 remained worse than both T1 and T2 after Holm correction (adjusted $p=.041$ for each). Mean top-ranked cosine similarity was .185 for T1, .320 for T2, and .174 for T3; T3 did not differ from T1 on this measure (adjusted $p=.833$) but was lower than T2 (adjusted $p<.001$). Among Top-3-hit tasks, median gold-versus-best-non-gold margins were .074, .176, and -.019 for T1, T2, and T3. The weakness therefore reflected reduced discrimination of the

correct card rather than uniformly low similarity. Because scores below rank 3 were not retained, exact margins for the 27 Top-3 misses could not be reconstructed; the analysis is interpreted as a post hoc mechanism diagnostic rather than a complete all-task margin estimate.

Linguistic Accuracy remained near the maximum in every condition. Run 1 means were 1.972 for B0, B1, and B2 (Friedman chi-square(2)=0.667, $p=.717$, Kendall's $W=.005$); Run 2 means were 1.958, 1.986, and 2.000 (chi-square(2)=2.000, $p=.368$, $W=.014$). These tests do not establish equivalence; they show no detected condition effect under a ceiling-limited benchmark (Table 3; Figure 3).

Retrieval Faithfulness likewise showed no reproducible B2 advantage. In Run 1, all 72 pairs were scorable: 2 improved from B1 to B2, 0 worsened, and 70 were tied (exact sign $p=.500$); means were 1.236 and 1.278 ($\text{delta}=+.042$). In Run 2, 70 pairs were scorable: 1 improved, 2 worsened, and 67 were tied (exact sign $p=1.000$); means were 1.314 and 1.300 ($\text{delta}=-.014$). Wilcoxon sensitivity tests were also non-significant ($p=.180$ and $.564$; Table 3). The sparse changes and reversal in descriptive direction provide no stable evidence of a B2 gain; without a prespecified equivalence margin, they also do not establish equivalence.

Task-level run-to-run stability was high for both primary outcomes. Linguistic Accuracy matched exactly in 211/216 episodes (97.69%) and within one score category in 213/216 (98.61%); its overall Spearman correlation was only $\rho=.273$ ($p<.001$) because the near-ceiling distribution offered little rank variance. Retrieval Faithfulness showed 139/142 exact matches (97.89%), 141/142 within-one matches (99.30%), and strong rank stability ($\rho=.975$, $p<.001$). These results support Run 2 as an operational task-level replication rather than merely a repetition of similar aggregate means; condition- and task-family breakdowns are reported in Supplementary Table S1.

The strongest replicated association was between retrieval rank and Retrieval Faithfulness. Spearman correlations ranged from approximately $-.83$ to $-.87$ across both runs and RAG conditions. When the gold card appeared in the top three, mean faithfulness was approximately 1.91-1.98; when it was missed, mean faithfulness was approximately .11-.16. This pattern was substantially more stable than the direct B1-B2 contrast (Figure 4).

The verifier showed little observable intervention: in each run, 71/72 B2 cases were ACCEPT and only P15-T1 was REVISE. In both runs, P15-T1 combined a REVISE decision with unchanged content. Run 1 also contained one ACCEPT-labelled candidate-final mismatch (P20-T2), but the archive cannot localize its source, so no mechanism is assigned. Overall, verifier decisions did not yield a stable measurable gain in final-response quality.

The two experts agreed on 231/234 applicable rating cells (98.72%), with 100% within-one agreement. Expert agreement was perfect for Linguistic Accuracy, Correction Restraint, Retrieval Faithfulness, and Unsupported Content; Pedagogical Appropriateness reached 98.15% exact agreement and Adaptivity 88.89%. Judge-human agreement was metric-specific: Linguistic Accuracy was complete, and Retrieval Faithfulness reached 94.44% exact agreement with each expert and weighted $\kappa=.904$. Correction Restraint and Pedagogical Appropriateness were ceiling- or zero-variance limited, while Unsupported Content and Adaptivity showed clear Judge-human limitations. Accordingly, Linguistic Accuracy and Retrieval Faithfulness were retained as primary automated outcomes; the remaining metrics were secondary or exploratory (Tables 4-5; Figure 5).

Table 2. Deterministic retrieval performance by task type.

Task type	N	Recall@1	Recall@3	MRR
Overall	72	.472	.625	.576
T1	24	.583	.708	.674
T2	24	.667	.792	.743
T3	24	.167	.375	.312

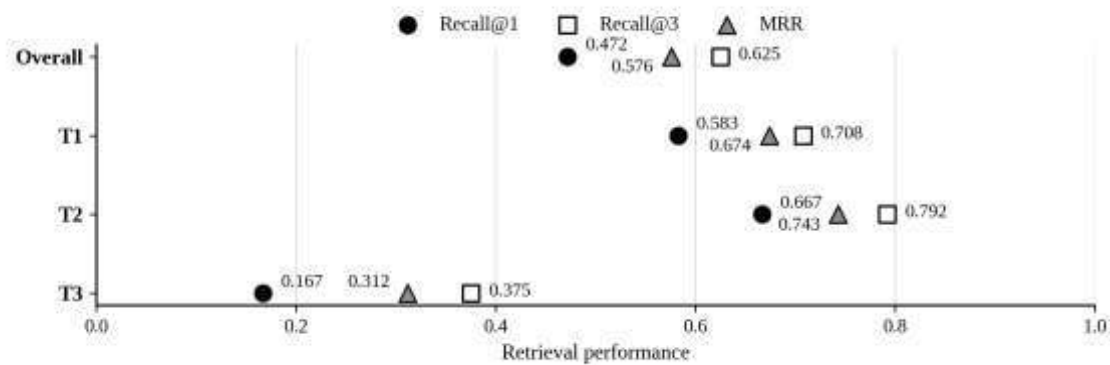


Figure 2. Deterministic retrieval performance by task type. Recall@1, Recall@3, and MRR are shown for all 72 tasks and for each 24-task family: T1, Diagnosis and Correction; T2, Explanation and Usage; and T3, Stateful Adaptive Tutoring. The frozen prompt-only TF-IDF retriever was identical across runs, so these values are the same in both runs.

Table 3. Primary outcomes across the two formal runs.

Outcome	Run	B0	B1	B2	Test	p
Linguistic Accuracy	1	1.972	1.972	1.972	Friedman	.717
Linguistic Accuracy	2	1.958	1.986	2.000	Friedman	.368
Retrieval Faithfulness	1	-	1.236	1.278	Wilcoxon (sens.)	.180
Retrieval Faithfulness	2	-	1.314	1.300	Wilcoxon (sens.)	.564

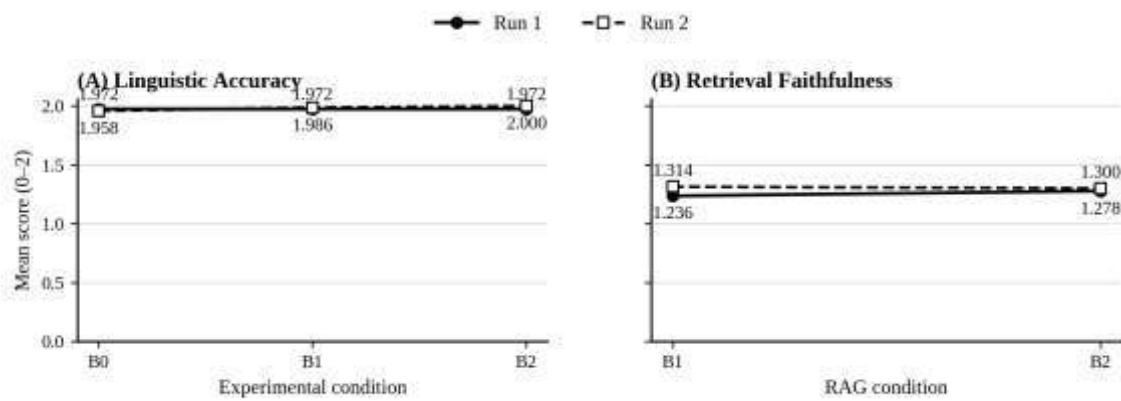


Figure 3. Primary automated outcomes across the two formal runs. (A) Mean Linguistic Accuracy (0-2) for B0, B1, and B2 ($n=72$ per condition/run): Run 1 Friedman chi-square(2)=0.667, $p=.717$; Run 2 chi-square(2)=2.000, $p=.368$. (B) Mean Retrieval Faithfulness (0-2) for B1 and B2: Run 1 $n=72$ paired, Wilcoxon $p=.180$, exact sign $p=.500$; Run 2 $n=70$ paired, Wilcoxon $p=.564$, exact sign $p=1.000$. Connecting lines are visual guides, not uncertainty intervals.

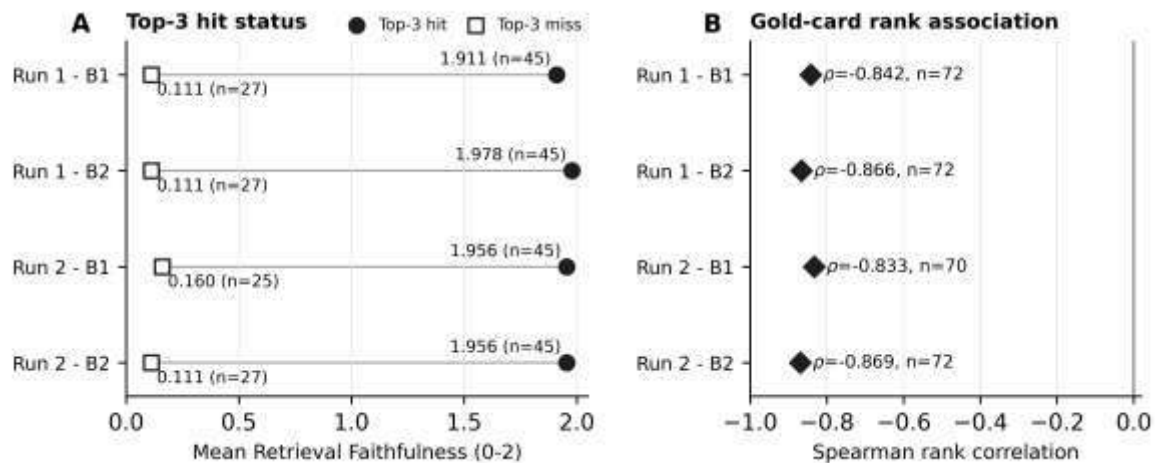


Figure 4. Retrieval quality and Retrieval Faithfulness across the two formal runs. (A) Mean Retrieval Faithfulness (0-2) for Top-3 gold-card hits versus misses; hit/miss counts were 45/27 (Run 1-B1), 45/27 (Run 1-B2), 45/25 (Run 2-B1; two non-applicable cases excluded), and 45/27 (Run 2-B2). (B) Spearman correlations between gold-card rank and Retrieval Faithfulness were $\rho=-.842$, $-.866$, $-.833$, and $-.869$, respectively (all $p<.001$). Lower numerical rank indicates better retrieval. Because faithfulness is defined relative to retrieved evidence, these associations do not establish a causal effect of rank on general response quality.

Table 4. Independent expert inter-rater reliability on the human-validation subset.

Metric	N	Exact agreement	Weighted kappa
Linguistic Accuracy	54	100%	1.000
Correction Restraint	18	100%	Not estimable
Retrieval Faithfulness	36	100%	1.000
Unsupported Content	54	100%	Not estimable
Pedagogical Appropriateness	54	98.15%	.658
Adaptivity	18	88.89%	.769

Table 5. Judge-human agreement on the stratified human-validation subset.

Metric	N	Judge-A exact	Judge-B exact	Weighted kappa (A / B)	Interpretation
Linguistic Accuracy	54	100.00%	100.00%	1.000 / 1.000	Primary automated outcome supported
Correction Restraint	18	100.00%	100.00%	NE / NE	Ceiling-limited; descriptive/secondary
Retrieval Faithfulness	36	94.44%	94.44%	0.904 / 0.904	Primary automated outcome supported
Unsupported Content	54	83.33%	83.33%	0.000 / 0.000	Exploratory; construct mismatch
Pedagogical Appropriateness	54	98.15%	96.30%	0.000 / 0.000	Secondary; Judge zero variance
Adaptivity	18	55.56%	66.67%	0.000 / 0.000	Exploratory; Judge ceiling bias

Note. Weighted kappa values are reported separately for Judge-Expert A and Judge-Expert B; NE = not estimable because all applicable ratings occupied one category. High exact agreement under zero variance does not establish discriminative validity.

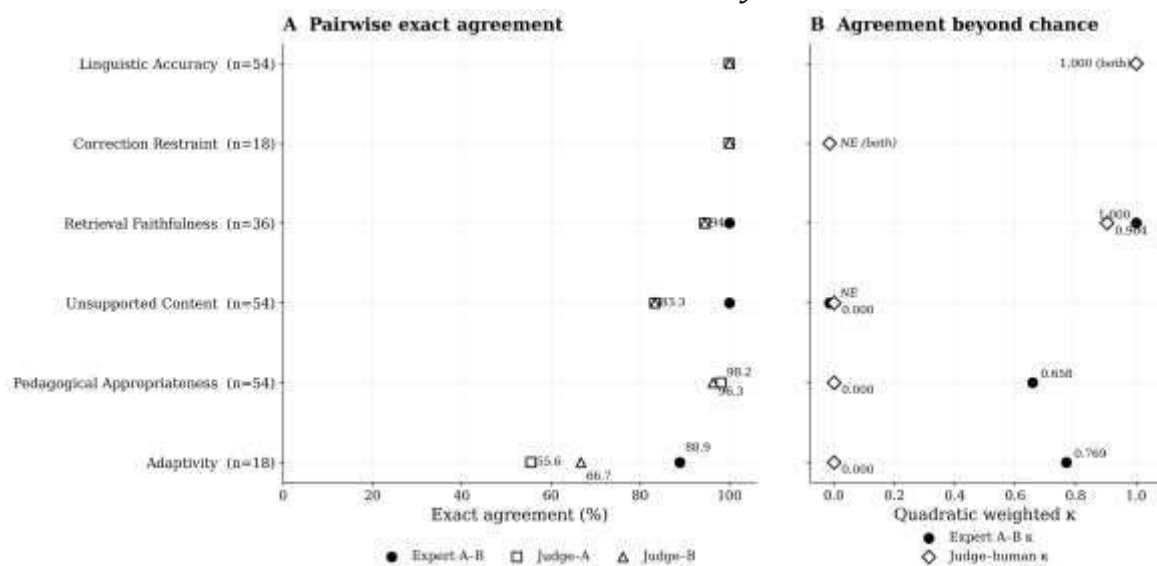


Figure 5. Human validation of automated evaluation metrics. (A) Pairwise exact agreement among Expert A, Expert B, and DeepSeek V4-Pro Judge V1.3 on the stratified 54-output subset. N varies by metric because Correction Restraint applies only to T1, Adaptivity only to T3, and Retrieval Faithfulness only to applicable RAG responses. (B) Quadratic weighted kappa; Judge-A and Judge-B values were identical within each metric, so one Judge-human series is shown. NE = not estimable because all applicable ratings occupied one category; high exact agreement under ceiling or zero variance does not imply discriminative validity.

5. Conclusion

This controlled two-run evaluation found near-ceiling Linguistic Accuracy with no reproducible condition effect, which does not establish equivalence. Retrieval Faithfulness showed no stable B2 advantage over B1. The strongest replicated result was that evidence

access tracked faithfulness: Top-3 gold-card hits scored high, misses much lower, and T3 retrieval discrimination was weaker. Because faithfulness is evidence-relative, this is a replicated pipeline dependency rather than a causal claim about overall response quality. The verifier rarely changed outputs and added no measurable benefit under the same-family, same-evidence design; learner-state effects remained unidentified because B2 bundled state conditioning and verification. Human validation supported Linguistic Accuracy and Retrieval Faithfulness more strongly than Adaptivity and Unsupported Content. Overall, decomposed evaluation of retrieval, grounding, verification, and metric validity is more informative than assuming added architectural components improve tutoring quality.

Research Transparency and Declarations

Data and materials. Retained materials comprise the frozen 72-task benchmark, 24 knowledge cards, prompts, retrieval, generated and verified outputs, automated-judge and human-validation records, token-use records, and available raw API records. Supplementary Table S2 distinguishes Run 2 identifiers directly recovered from the archive from historical Run 1 identifiers lacking per-case payloads. Credential-scrubbed materials are available to the editorial office on reasonable request.

Ethical and expert-rater boundary. The study used synthetic tutoring tasks and AI-generated outputs; no learners were recruited and no student records, personal learner data, or learning outcomes were collected. Two linguistic experts rated anonymized outputs and task materials, consented to use of their anonymized ratings, and had no personal characteristics analyzed.

AI systems and manuscript assistance. DeepSeek V4-Flash and DeepSeek V4-Pro served as generation, verification, and automated-evaluation components described in Materials and Methodology; their outputs were treated as experimental system outputs. ChatGPT assisted only with language polishing and document formatting; the authors retained responsibility for scientific content, statistical interpretation, reference verification, and the final manuscript.

Author contributions. Shanning Fu led conceptualization, methodology, experimental implementation, data curation, formal analysis, visualization, and manuscript preparation and revision. Ya Han supported validation, data review, and manuscript review. Dilshod Bekmuratov provided manuscript checking and supervisory quality control. All authors approved the final manuscript and submission.

Originality, funding, and conflicts. The manuscript is original, is not under consideration elsewhere, and is approved by all authors for submission to Topical Issues of Technical Sciences (Techscience.uz). This research received no external funding, and the authors declare no conflicts of interest.

Adabiyotlar/Литература/References:

1. Adenuga J.A. Retrieval-augmented generation for pedagogically aware educational AI: an expert-rated comparison of a prompt-only LLM tutor and an integrated, learner-state-aware RAG tutor. *Frontiers in Education*, 2026, Vol. 11, 1896839. <https://doi.org/10.3389/feduc.2026.1896839>
2. Chen G.H., Chen S., Liu Z., Jiang F., Wang B. Humans or LLMs as the Judge? A Study on Judgement Bias. *Proceedings of EMNLP 2024*, 2024, pp. 8301-8327. <https://doi.org/10.18653/v1/2024.emnlp-main.474>
3. Es S., James J., Espinosa Anke L., Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of EACL 2024: System Demonstrations*, 2024, pp. 150-158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>

4. Fang F., Bai Y., Ni S., Yang M., Chen X., Xu R. Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training. *Proceedings of ACL 2024*, 2024, pp. 10028-10039. <https://doi.org/10.18653/v1/2024.acl-long.540>
5. Gou Z., Shao Z., Gong Y., et al. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html
6. Hattie J., Timperley H. The Power of Feedback. *Review of Educational Research*, 2007, Vol. 77, No. 1, pp. 81-112. <https://doi.org/10.3102/003465430298487>
7. Jain M., Bhatt T., Pitts G., Pandya A., Brusilovsky P., Norouzi N., Hellas A., Leinonen J., Akram B. Retrieval-Augmented Tutoring for Algorithm Tracing and Problem-Solving in AI Education. *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, 2026, pp. 841-850. <https://doi.org/10.18653/v1/2026.bea-1.57>
8. Kamoi R., Zhang Y., Zhang N., Han J., Zhang R. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. *Transactions of the Association for Computational Linguistics*, 2024, Vol. 12, pp. 1417-1440. <https://doi.org/10.1162/tacl.a.00713>
9. Kohnke L., Moorhouse B.L., Zou D. ChatGPT for Language Teaching and Learning. *RELC Journal*, 2023, Vol. 54, No. 2, pp. 537-550. <https://doi.org/10.1177/00336882231162868>
10. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 2020, Vol. 33, pp. 9459-9474. https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
11. Li Y., Zhou K., Qiao Q., Nguyen B., Wang Q., Li Q. Investigating Context Faithfulness in Large Language Models: The Roles of Memory Strength and Evidence Style. *Findings of ACL 2025*, 2025, pp. 4789-4807. <https://doi.org/10.18653/v1/2025.findings-acl.247>
12. Liu Y., Iter D., Xu Y., Wang S., Xu R., Zhu C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *Proceedings of EMNLP 2023*, 2023, pp. 2511-2522. <https://doi.org/10.18653/v1/2023.emnlp-main.153>
13. Lyster R., Ranta L. Corrective Feedback and Learner Uptake: Negotiation of Form in Communicative Classrooms. *Studies in Second Language Acquisition*, 1997, Vol. 19, No. 1, pp. 37-66. <https://doi.org/10.1017/S0272263197001034>
14. Ma W., Adesope O.O., Nesbit J.C., Liu Q. Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of Educational Psychology*, 2014, Vol. 106, No. 4, pp. 901-918. <https://doi.org/10.1037/a0037123>
15. Madaan A., Tandon N., Gupta P., et al. Self-Refine: Iterative Refinement with Self-Feedback. *Advances in Neural Information Processing Systems*, 2023, Vol. 36, pp. 46534-46594. <https://doi.org/10.52202/075280-2019>
16. Nguyen V.D., Tran N.H., Dao K.A., Nguyen P.V. Towards Personalized AI Education: Context-Aware Retrieval-Augmented Generation With Grade-Level LLM Adaptation. *Computer Applications in Engineering Education*, 2026, Vol. 34, No. 1, e70153. <https://doi.org/10.1002/cae.70153>
17. Niu C., Wu Y., Zhu J., et al. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. *Proceedings of ACL 2024*, 2024, pp. 10862-10878. <https://doi.org/10.18653/v1/2024.acl-long.585>
18. Pan L., Saxon M., Xu W., Nathani D., Wang X., Wang W.Y. Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies. *Transactions of the Association for Computational Linguistics*, 2024, Vol. 12, pp. 484-506. <https://doi.org/10.1162/tacl.a.00660>
19. Park S.-I., Lee J.-Y. Toward Robust RALMs: Revealing the Impact of Imperfect Retrieval on Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 2024, Vol. 12, pp. 1686-1702. <https://doi.org/10.1162/tacl.a.00724>
20. Saad-Falcon J., Khattab O., Potts C., Zaharia M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *Proceedings of NAACL 2024*, 2024, pp. 338-354. <https://doi.org/10.18653/v1/2024.naacl-long.20>
21. Shi H., Aryadoust V. A systematic review of AI-based automated written feedback research. *ReCALL*, 2024, Vol. 36, No. 2, pp. 187-209. <https://doi.org/10.1017/S0958344023000265>
22. VanLehn K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other

- Tutoring Systems. Educational Psychologist, 2011, Vol. 46, No. 4, pp. 197-221. <https://doi.org/10.1080/00461520.2011.611369>
23. Wang P., Li L., Chen L., et al. Large Language Models are not Fair Evaluators. Proceedings of ACL 2024, 2024, pp. 9440-9450. <https://doi.org/10.18653/v1/2024.acl-long.511>
 24. Wang X., Maeda Y., Chang H.-H. Development and techniques in learner model in adaptive e-learning system: A systematic review. Computers & Education, 2025, Vol. 225, 105184. <https://doi.org/10.1016/j.compedu.2024.105184>
 25. Wang X., Wang Z., Gao X., et al. Searching for Best Practices in Retrieval-Augmented Generation. Proceedings of EMNLP 2024, 2024, pp. 17716-17736. <https://doi.org/10.18653/v1/2024.emnlp-main.981>
 26. Yan L., Greiff S., Teuber Z., Gašević D. Promises and challenges of generative artificial intelligence for human learning. Nature Human Behaviour, 2024, Vol. 8, pp. 1839-1850. <https://doi.org/10.1038/s41562-024-02004-5>
 27. Yang L., Li R. ChatGPT for L2 learning: Current status and implications. System, 2024, Vol. 124, 103351. <https://doi.org/10.1016/j.system.2024.103351>
 28. Zhang Q., Xiang Z., Xiao Y., Wang L., Li J., Wang X., Su J. FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation. Proceedings of ACL 2025, 2025, pp. 21863-21882. <https://doi.org/10.18653/v1/2025.acl-long.1062>

TECHSCIENCE.UZ

**TEXNIKA FANLARINING DOLZARB
MASALALARI**

№ S/4 (4)-2026

TOPICAL ISSUES OF TECHNICAL SCIENCES

**TECHSCIENCE.UZ- TEXNIKA
FANLARINING DOLZARB MASALALARI**
elektron jurnali 15.09.2023-yilda 130346-
sonli guvohnoma bilan davlat ro'yxatidan
o'tkazilgan.

Muassislar: "SCIENCEPROBLEMS TEAM"
mas'uliyati cheklangan jamiyati;
Jizzax politexnika insituti.

TAHRIRIYAT MANZILI:

Toshkent shahri, Yakkasaroy tumani, Kichik
Beshyog'och ko'chasi, 70/10-uy.

Elektron manzil:

scienceproblems.uz@gmail.com